

# Phylogenetically and Spatially Conserved Word Pairs Associated with Gene Expression Changes in Yeasts

**UC Berkeley**

Derek Chiang  
Alan Moses  
Mike Eisen

**MIT**

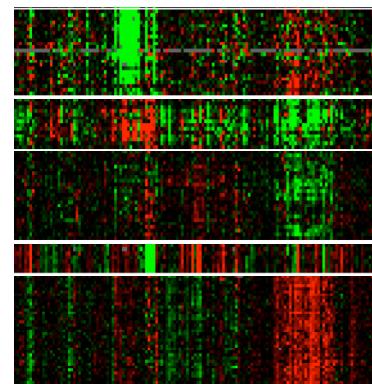
Manolis Kamysselis  
Eric Lander

## Regulation of Gene Expression

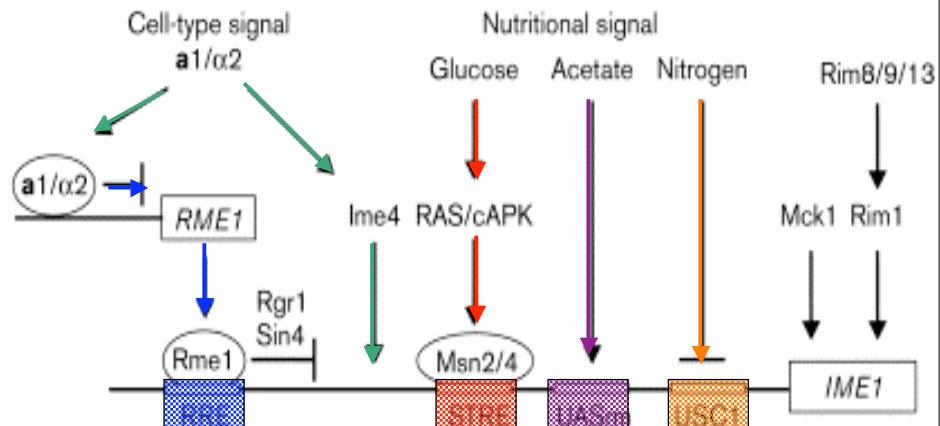
How is information regulating gene expression encoded in genome sequences?

```
>Saccharomyces cerevisiae chr V
CGTCTCTCCAAGCCCTGGTGTCTTACCC
GGATGTTCAACAAAAGCTACTTACTACCTT
TATTTTATGTTACTTTTATAGATTGTCTT
TTTATCTTACTCTTCCCACTTGTCCTCGC
TACTGCCGTGCAACAAACACTAAATCAAAC
AGTGAAATACTACTACATCAAACGCATATT
CCCTAGAAAAAAATTCTTACAATATACT
ATACTACACAATACATAATCACTGACTTTCG
TAACAAACAATTCCCTCACTCTCAACTCT
CTGCTCGAATCTCTACATAGTAATATTATAT
CAAATCTACCGTCTGGAACATCATCGCTATC
CAGCTTTGTGAACCGCTACCATCAGCATG
TACAGTGGTACCTTCGTGTTATCTGCAGCGA
GAACCTCAACGTTGCCAAATCAAGCCAATG
TGGTAACAAACCACCTCCGAATCTGCTCC
AAAAGATACTCCAGTTCTGCCGAATGTTT
```

Features?

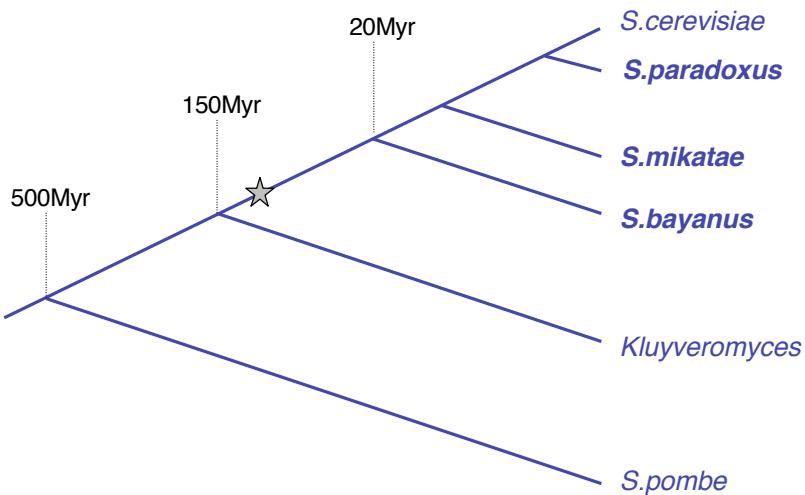


## Multifactorial Gene Regulation



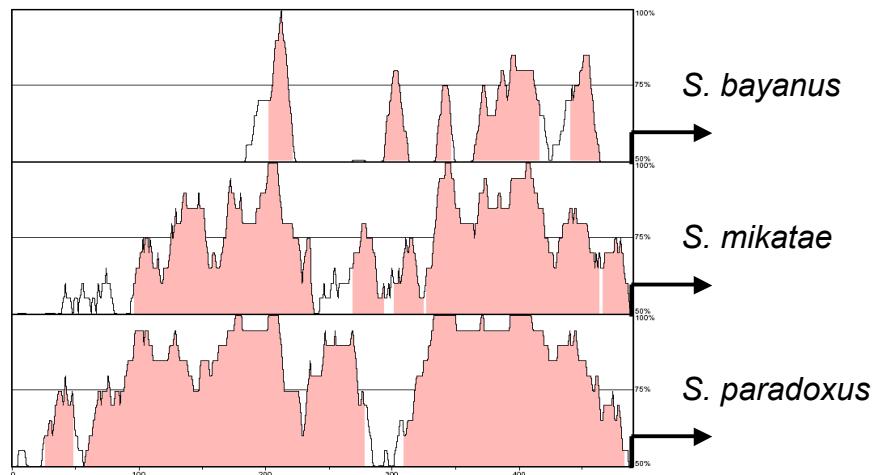
REF: Verson, A. K. & Pierce, M. (2000)  
*Curr. Opin. Cell Biol.* **12**:334-339

## Comparative Genomics



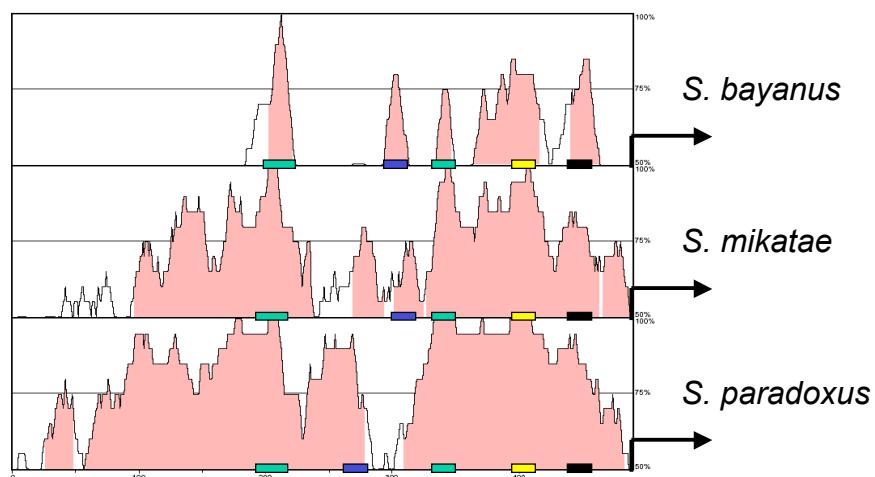
## Phylogenetic Footprinting

*MET28* upstream region

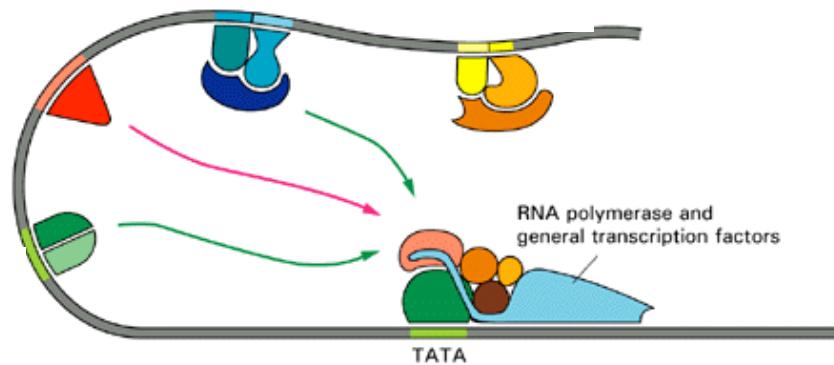


## Phylogenetic Footprinting

*MET28* upstream region

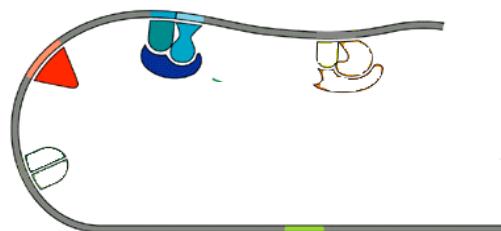


## From Footprints to Promoter Structure



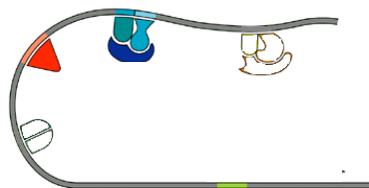
REF: Alberts B, et al (1993).  
Molecular Biology of the Cell

## Conserved Word Pair Templates



- 1) Joint Conservation
- 2) Close Spacing
- 3) Validate with Gene Expression

## TEST #1: Joint Word Conservation



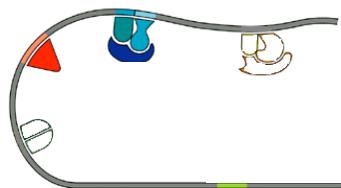
Word 1  
Conserved  
Y  
N

		Word 2 Conserved	
		Y	N
Y	Y		
	N		

3860 intergenic regions (CLUSTALW)

Conserved = Identical in 3+ genomes  
within 600 bp of gene start

## TEST #1: Joint Word Conservation



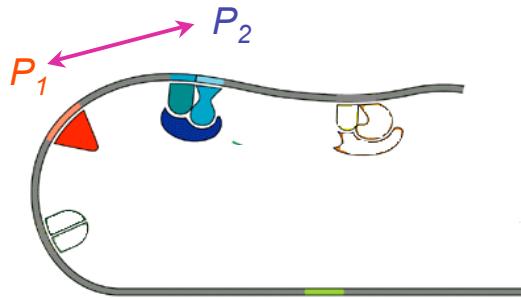
Word 1  
Conserved  
Y  
N

		Word 2 Conserved	
		Y	N
Y	Y	32	162
	N	134	3226

**Chi-square Test for Independence**  
(Yates adjustment)

$$\chi^2 = \sum_k \frac{(|O_k - E_k| - \frac{1}{2})^2}{E_k}$$

## Conserved Word Pair Templates

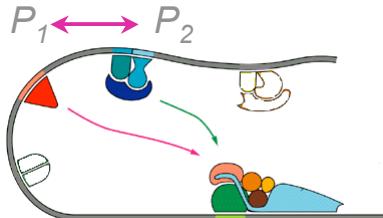


1) Joint Conservation (8452 pairs)

### 2) Close Spacing

3) Validate with Gene Expression

## TEST #2: Close Spacing



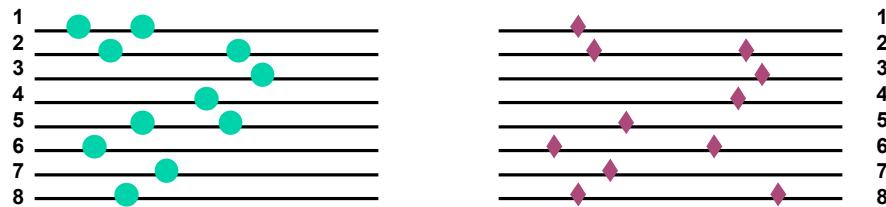
Average minimum distance

$$\overline{D} = \frac{1}{N} \sum_g \min_{k \in g_i} |P_{1k} - P_{2k}|$$

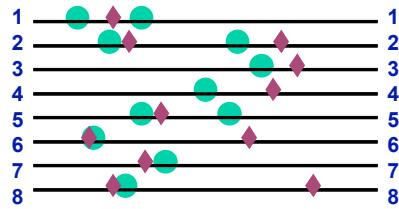
- Use positions of conserved sites in *S. cer*
- Evaluate significance of  $\overline{D}$  using an empirical null distribution (Permutation test)

## TEST #2: Close Spacing

### Permutation Test

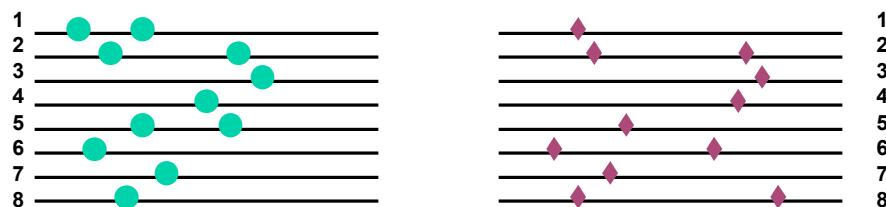


Avg. Min. Distance  
 $D = 32 \text{ bp}$

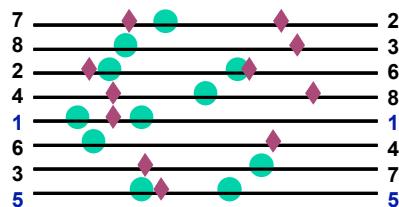


## TEST #2: Close Spacing

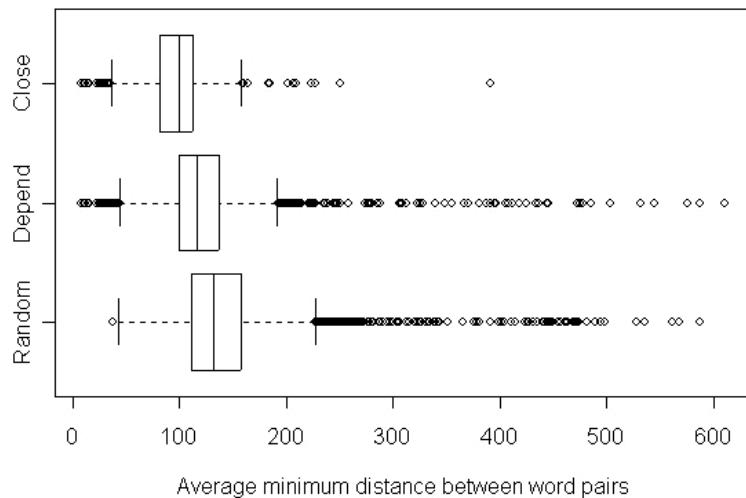
### Permutation Test



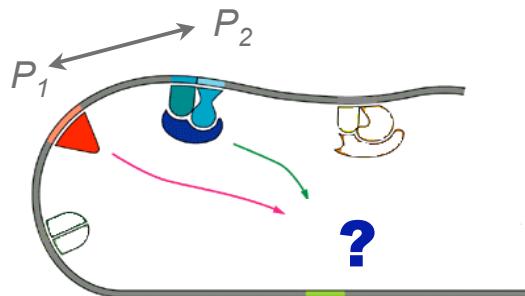
Permutation Distance  
 $D^* = 65 \text{ bp}$



## TEST #2: Close Spacing



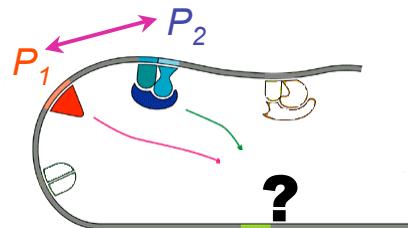
## Conserved Word Pair Templates



- 1) Joint Conservation (8452 pairs)
- 2) Close Spacing (1117 pairs)
- 3) Validate with Gene Expression**

## Validating Expression Subsets

### Group by Sequence Approach



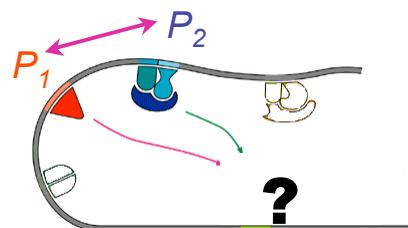
Genome (5500 genes)

{  $P_1$  &  $P_2$  conserved  
and closely spaced }

SUBSET ( $N$  genes)

## Validating Expression Subsets

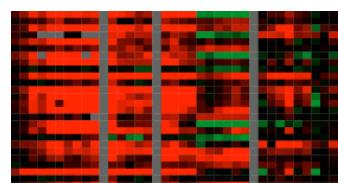
### Group by Sequence Approach



Genome (5500 genes)

{  $P_1$  &  $P_2$  conserved  
and closely spaced }

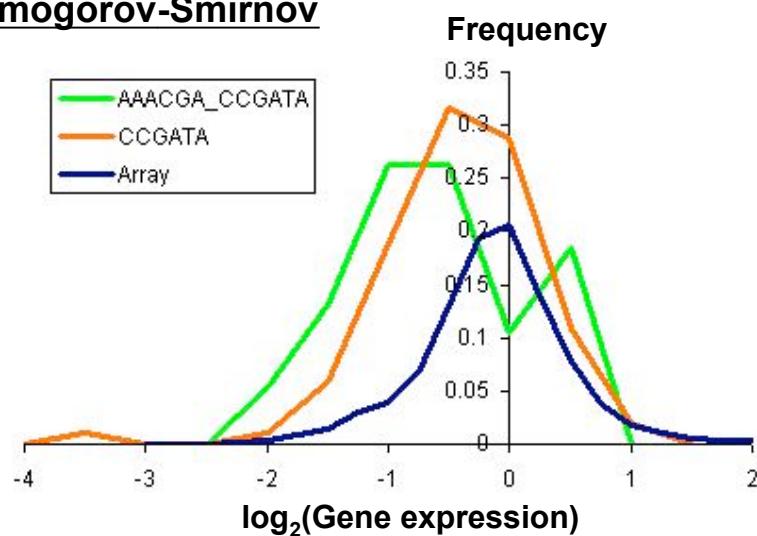
SUBSET ( $N$  genes)



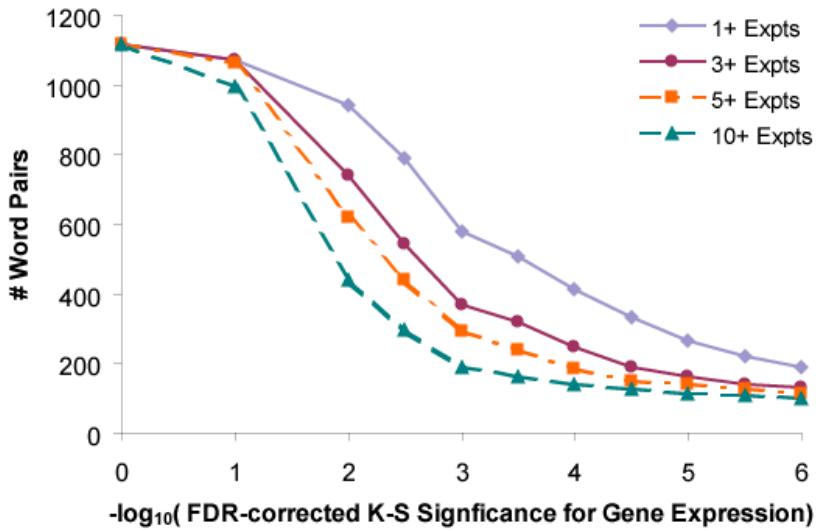
Associate with gene expression

## Validating Expression Subsets

### Kolmogorov-Smirnov

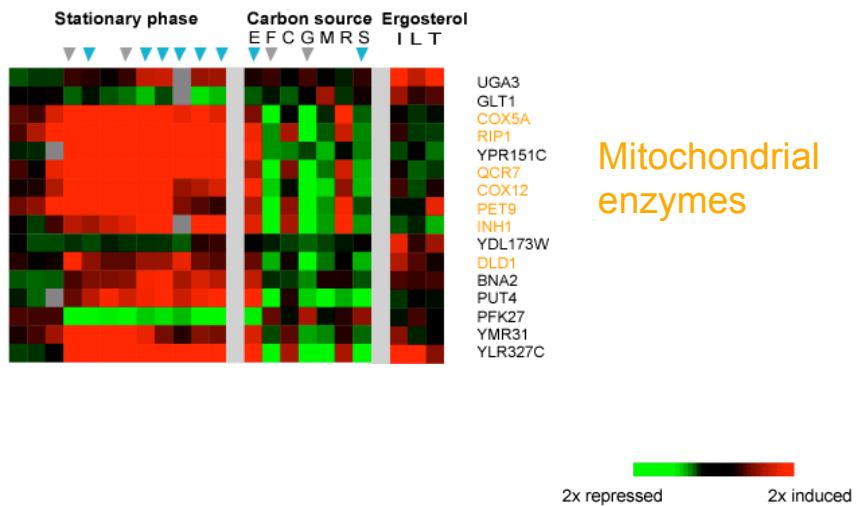


## Validating Expression Subsets



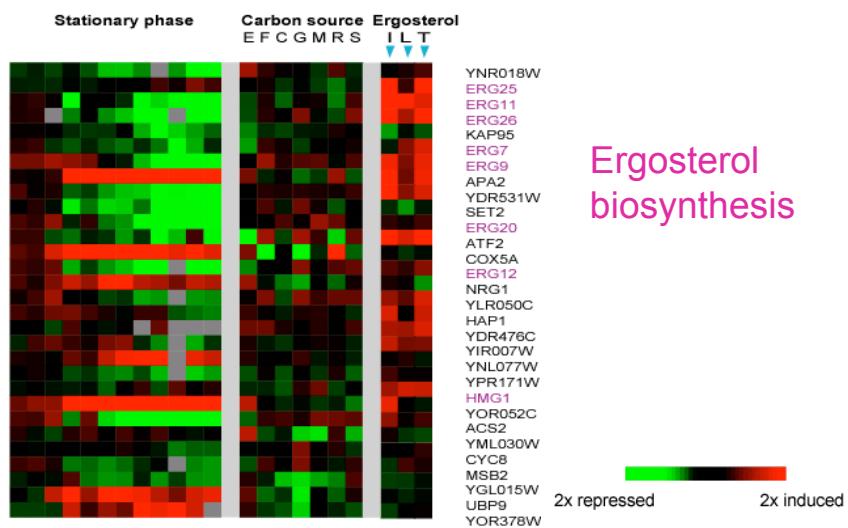
## More Informative Predictors

### CCGATA-CCAATC (Hap1-Hap4)

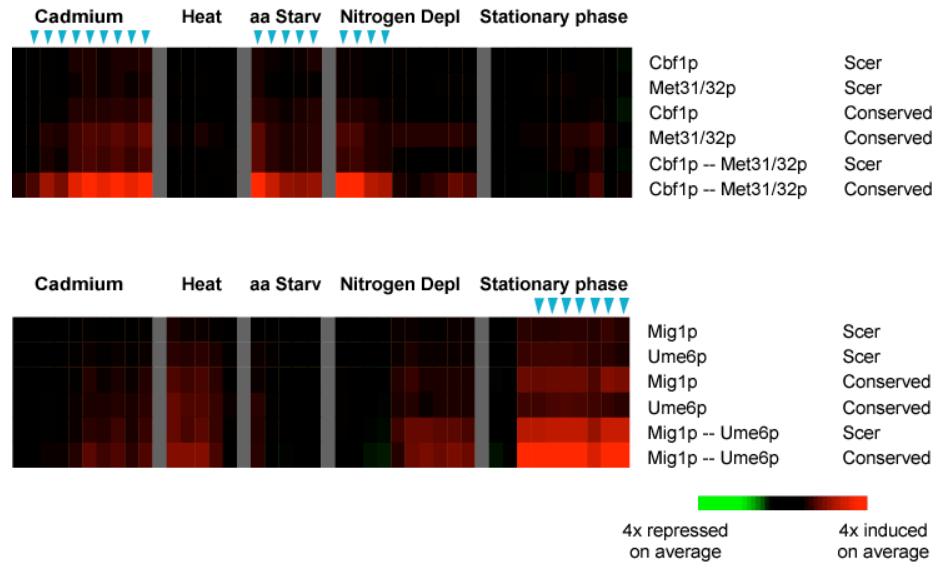


## More Informative Predictors

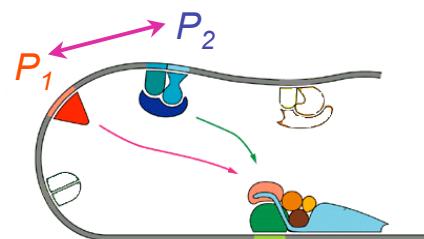
### CCGATA-TCGTTT (Hap1-Upc2)



## More Informative Predictors

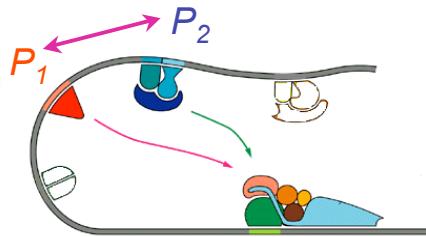


## Summary



- Conservation of promoter structure yields highly specific predictors of gene expression
- Significant gene expression changes in conditions where TF's are known to be active

## Future Directions



- Use probabilistic sequence & distance models
- Prove biological mechanism with experiments

## Public Library of Science



<http://www.publiclibraryofscience.org>

- Free & unfettered access to full text articles
- PLoS Journals coming this fall
- Editorial Staff
  - Vivian Siegel (former Senior Editor, *Cell*)
  - Mark Patterson (former Editor, *Nature Rev Genetics*)
  - Phil Bernstein (former Editor, *J Exper Medicine*)
  - Barbara Cohen (former Editor, *J Clinical Investigation*)

## Acknowledgements

### UC Berkeley

Peter Bickel  
John Storey  
Mark van der Laan

### Funding

US Dept of Energy  
Pew Institute  
HHMI

### Eisen Lab

Justin Fay  
Audrey Gasch  
Hunter Fraser  
Venky Nandagopal  
Dan Pollard